Weekly Report

Yuxin Ma

08.17.2015 - 08.23.2015

Projects

Spark Workflow System

1. Current Progress

A basic workflow framework has been implemented by Zhendong Cao, including:

- Data importing;
- Workflow design;
- Workflow execution.

2. Discussion with Prof. Chen and Wu on Wednesday

Summary of the minute:

- First we should focus on a specific algorithm. The clustering task is a good way to start.
- Then we need to summarize what types of monitoring information can be used in Spark.
- The performance data derived from Spark is to be combined with the intermediate output of the selected algorithm. The relation between the intermediate result and the performance data will be studied.
- Basic visual design principles are to be added based on the available data.

3. Survey

Databricks has published their visualization tool for monitoring the execution progress of Spark system[1]. There are two main components that are related to our project:

- **Spark Events Timeline Visualization** (Figure 1) The process of `map` and `reduce` contains several running events marking the execution status. The event sequence is visualized as Gantt chart in the timeline.
- **Execution DAG Visualization** (Figure 2) The DAG graph is dedicated to represent the data reuse behavior of the process. The frequently re-used data should be considered to be cached in RAM.

[1] https://databricks.com/blog/2015/06/22/understanding-your-spark-application-through-visualization.html

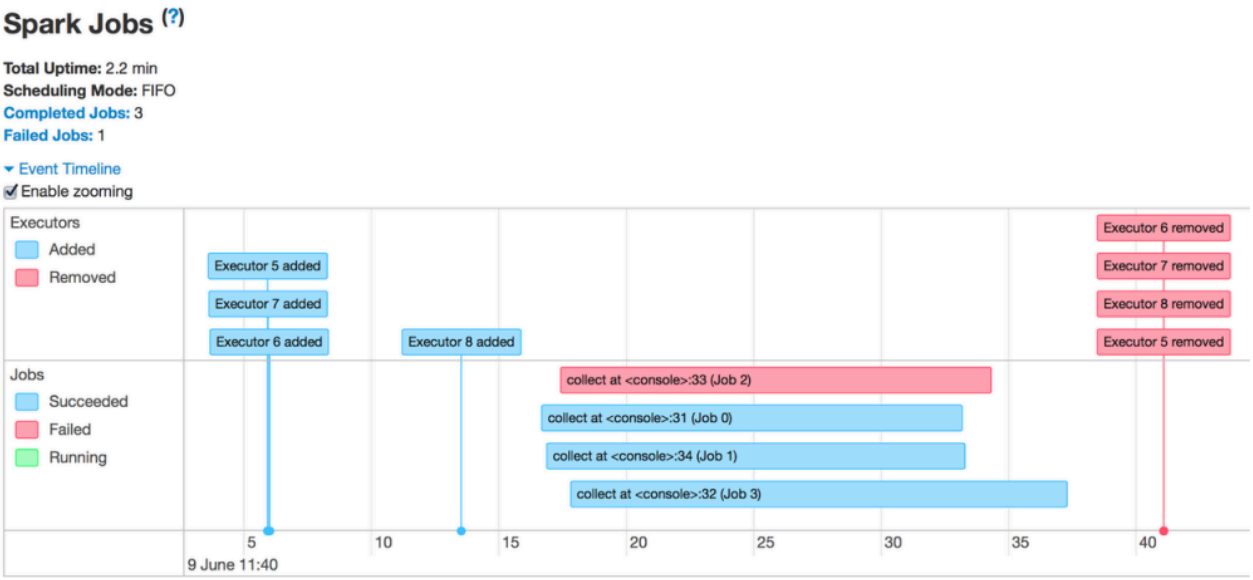

Figure 1: Timeline view of execution events. Each bar represents the start and finish time of the corresponding event.

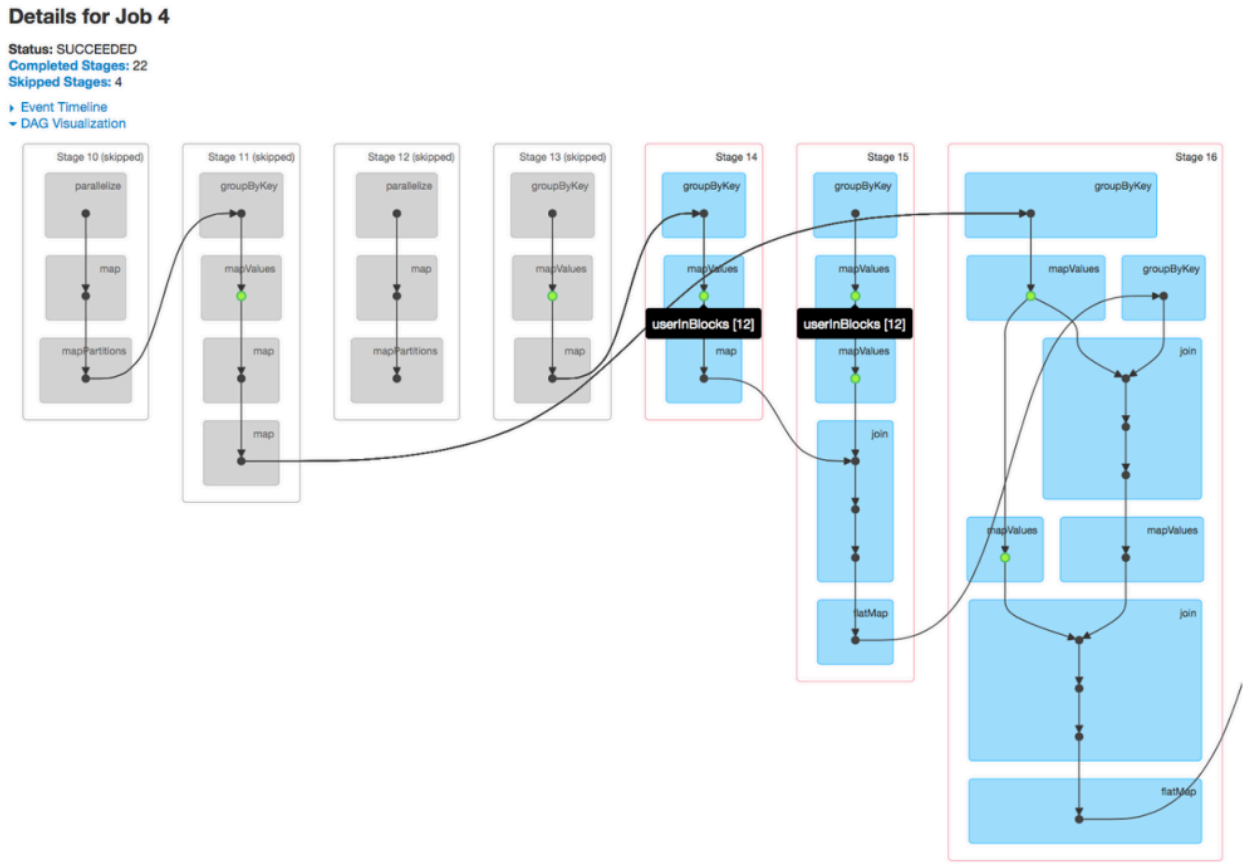

Figure 2: Execution DAG Graph. For complex procedure with multiple map-reduce tasks, the arrow indicates the data reuse behavior among different map-reduce processes.

We may extract the execution data from the visualization tool described in 3. Besides the visual design provided by Databricks, we can have our own design based on both execution data and the intermediate output.

Besides, Microsoft released their interactive machine learning platform[2] (Figure 3). It has a visual programming system for users to design their analysis workflow. The components cover from data pre-processing, machine learning and basic statistical charts.

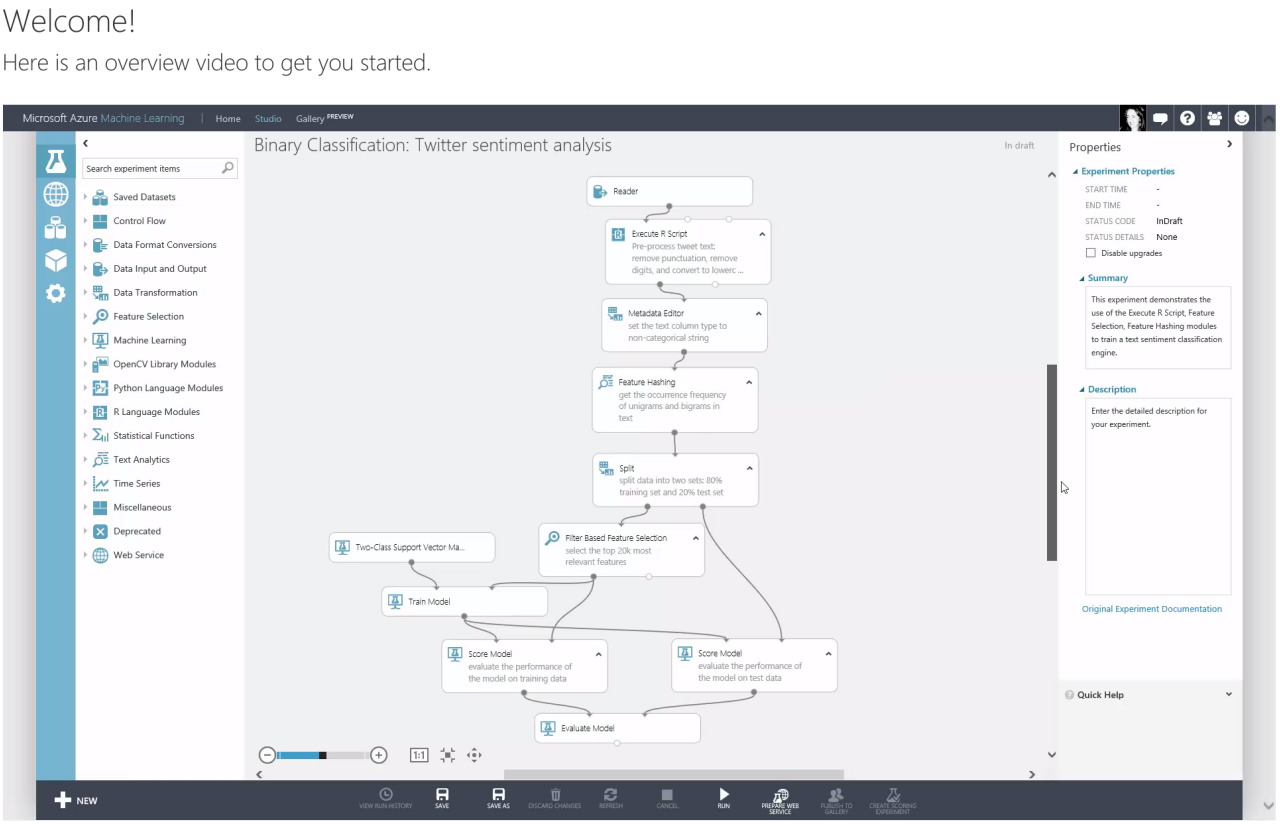

Figure 3: Microsoft Azure Machine Learning Platform.

4. Dataset

Currently there are three dataset available:

- **User attributes in Twitter.** Similar to [1], the attributes can be extracted from the raw Twitter data.
- **Trajectories from Call Detail Records.** However a quick look of trajectory clustering methods is needed

Huawei Large Graph Project

Software Design Specification

The sections of architecture design and system design have been delivered to Fangzhou. Currently there are two system plans for the visualization server:

[2]https://studio.azureml.net

1. Python Flask (web framework) + C modules (layout algorithms), which we are familiar with;

2. Golang Martini (web framework) + Golang modules (layout algorithms), which is of high execution performance.

After discussion we decided to use the Python way because of the learning cost.

Plan for the Next Week

For Spark Workflow System:

- Take a quick look at the trajectory clustering methods to see whether k-means can be applied, or there are other iterative clustering methods we can utilize.
- Zhendong is going to investigate the visualization tool to find out how to access the execution data.
- Write a draft including motivation, contribution, possible solutions and expected case study result.

For Huawei Large Graph Project:

- Create a basic Python Flask framework for Huawei Large Graph Project so others can start implementing server-side components.

References

[1] Z. Wang, C. Chen, J. Zhou, J. Liao, W. Chen, and R. Maciejewski, "A novel visual analytics approach for clustering large-scale social data," in *Big Data, 2013 IEEE International Conference on*, pp. 79–86, Oct 2013.